A method extracts an alpha matte from a video acquired of a scene. A set of pinhole or narrow aperture images $I_p$ is acquired of the scene with a camera aperture set to a relatively large depth-of-field. The scene includes a background $B$ and a foreground $F$. A corresponding set of wide aperture images $I_w$ is acquired of the scene with the camera aperture set to a relatively small depth-of-field. The respective pinhole and wide aperture images are combined to extract an alpha matte according to

$$\alpha = 1 + (I_p - I_w)(B - B_p).$$

19 Claims, 11 Drawing Sheets
OTHER PUBLICATIONS


* cited by examiner
\[ I_P = \alpha F + (1 - \alpha)B \]
\[ I_F = \alpha F + (1 - \alpha)B_F \]
\[ \frac{1 + (I_F - I_P)}{(B - B_F)} \]
SYSTEM AND METHOD FOR DEFOCUS DIFFERENCE MATING

FIELD OF THE INVENTION

This invention relates generally to image editing, and more particularly to matting.

BACKGROUND OF THE INVENTION

Matting and compositing are frequently used in image and video editing, 3D photography, and film production. Matting separates a foreground region from an input image by estimating a color $F$ and an opacity $a$ for each pixel in the image. Compositing uses the matte to blend the extracted foreground with a novel background to produce an output image representing a novel scene. The opacity $a$ measures a "coverage" of the foreground region, due to either partial spatial coverage or partial temporal coverage, i.e., motion blur. The set of all opacity values $a$ is called the alpha matte, the alpha channel, or simply a matte.

The matting problem can be formulated as follows. An image of a foreground against an opaque black background in a scene is $aF$. An image of the background without the foreground is $B$. An alpha image, where each pixel represents a partial coverage of that pixel by the foreground, is $a$. The image $a$ is essentially an image of the foreground object "painted" white, evenly lit, and held against the opaque background. The scale and resolution of the foreground and background images can differ due to perspective shortening.

The notions of an alpha matte, pre-multiplied alpha, and the algebra of composition have been formalized by Porter et al., "Compositing digital images," in Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press, pp. 253-259, 1984. They showed that for a pinhole (narrow aperture) camera, the image $aF$ in front of the background image $B$ can be expressed $501$ (see FIG. 5) by a linear interpolation:

$$I_p = aF + (1-a)B$$

where $I_p$ is a pinhole (narrow aperture) image, $aF$ is the pre-multiplied image of the foreground against an opaque background, and $B$ is the image of the opaque background in the absence of the foreground.

Matting is the inverse problem of solving for the unknown values of the variables ($a$, $F_x$, $F_y$, $F_z$, $B_x$, $B_y$, $B_z$), given the composite image pixel values ($I_p$, $I_x$, $I_y$), where $x$, $y$, and $z$ are color channels. The "P" subscript denotes that Equation (1) holds for a pinhole camera, i.e., where the entire scene is in focus. One can approximate a pinhole camera with a very narrow aperture. Blue screen matting is easier to solve because the background color $B$ is known.


Conventional matting requires a background with known, constant color, which is referred to as blue screen matting. If a digital camera is used, then a green matte is preferred. Blue screen matting is the predominant technique in the film and broadcast industry. For example, broadcast studios use blue matting for presenting weather reports. The background is a blue screen, and the foreground region includes the weatherman standing in front of the blue screen. The foreground is extracted, and then superimposed onto a weather map so that it appears that the weatherman is actually standing in front of the map. However, blue screen matting is costly and not readily available to casual users. Even production studios would prefer a lower-cost and less intrusive alternative.


It desired to perform matting using non-intrusive techniques. That is, the scene does not need to be modified. It is also desired to perform the matting automatically. Furthermore, it is desired to provide matting for "rich" natural images, i.e., images with a lot of fine, detailed structure.

Most natural image matting methods require manually defined trinaps to determine the distribution of color in the foreground and background regions. A trinap segments an image into background, foreground and unknown pixels. Using the trinaps, those methods estimate likely values of the foreground and background colors of unknown pixels, and use the colors to solve the matting Equation (1).

Bayesian matting, and its extension to image sequences, produces the best results in many applications. However, those methods require manually defined trinaps for key frames. This is tedious for a long image sequence. It is desired to provide a method that does not require user intervention, and that can operate in real-time as an image sequence is acquired.

The prior art estimation of the color distributions works only when the foreground and background are sufficiently different in a neighborhood of an unknown pixel. Therefore, it is desired to provide a method that can extract a matte where the foreground and background pixels have substantially similar color distributions.

The Poisson matting of Sun et al. solves a Poisson equation for the matte by assuming that the foreground and background are slowly varying. Their method interacts closely with the user by beginning from a manually constructed trinap. They also provide "painting" tools to correct errors in the matte.

An unassisted, natural video matting system is described by Zitnick et al., "High-quality video view interpolation using a layered representation," ACM Trans. on Graphics 23, 3, pp. 600-608, 2004. They acquire videos with a horizontal row of eight cameras spaced over about two meters. They measure depth discrepancies from stereo disparity using sophisticated region processing, and then construct a trinap from the depth
3 discrepancies. The actual matting is determined by the Baye-

sian matting of Chuang et al. However, that method has the

view dependent problems that are unavoidable with stereo

cameras, e.g., reflections, specular highlights, and occlu-
sions. It is desired to avoid view dependent problems.

Difference matting, also known as background subtraction,
solves for $\alpha$ and the alpha-multiplied foreground $\alpha F$

given background and trimap images, Qian et al., “Video

camera background replacement without a blue screen,” Pro-
ceedings of ICIP, vol. 4, 143-146, 1999. However, difference matting

has limited discrimination at the borders of the foreground.

Another method uses back lighting to determine the matte.

Back lighting is a common segmentation method used in

many computer vision systems. Back lighting has also been

used in image-based rendering systems, Debevec et al., “A

Combining approaches to live action compositing,”


That method has two drawbacks. First, active illumination

is required, and second, incorrect results may be produced near

object boundaries because some objects become highly

reflective near grazing angles of the light.

Scene reconstruction is described by Favaro et al., “See-

ing beyond occlusions (and other marvels of a finite lens

aperture),” Proc. of the IEEE Intl. Conf. on Computer Vision

and Pattern Recognition, p. 579, 2003. That method uses de-

focal images and gradient descent minimization of a sum-

squared error. The method solves for coarse depth and a

binary alpha.

Another method uses a depth-from-focus system to

recover overlapping objects with fractional alphas, Sche-

chner et al., “Separation of transparent layers using focus,”


They position a motorized CCD axially behind a lens to acquire

images with slightly varying points of focus. Depth is recov-

ered by selecting the image plane location that has the best

focussed image. That method is limited to static scenes.

Another method uses three video streams acquired by three

cameras with different depth-of-field and focus and that share

the same center of projection to extract mattes for scenes with

unconstrained, dynamic backgrounds, McGuire et al., “De-

focal Video Matting,” ACM Transactions on Graphics 24, 3,

2003; and U.S. patent application Ser. No. 11/092,376, filed

by McGuire et al. on Mar. 29, 2005, “System and Method for

Image Matting.”

SUMMARY OF THE INVENTION

Matting is a process for extracting a high-quality alpha

matte and foreground from an image or a video. Conventional

techniques require either a known background, e.g., a blue

screen, or extensive manual interaction, e.g., manually speci-

fied foreground and background regions. Matting is generally

under-constrained, because not enough information is

obtained when the images are acquired.

One embodiment of the invention provides a system

and method for extracting a matte automatically from a video.

The video includes sets of pinhole (narrow aperture) images

and wide apertures images (frames) that are produced either

in parallel or in a time-interleaved manner.

The parallel sets of images can be acquired with a camera

having two optical systems that have a common center of

projection. One optical system has a large depth-of-field to

acquire the pinhole images, while the other optical system has

a small depth-of-field to acquire the wide apertures images.

A single camera can acquire the time-interleaved images

using a fast switching aperture. The aperture includes polar-

izing elements that can rapidly switch between different aper-

ture sizes. As an advantage, the aperture size is manipulated

using optical techniques. Thus, the aperture does not require

any moving parts, and can be switched at rates far exceeding

mechanical apertures.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a method for extracting a matte

from a video according to an embodiment of the invention;

FIG. 2 is a block diagram of a method for extracting a matte

from a video according to an embodiment of the invention;

FIG. 3A is an exploded diagram of a camera aperture

according to an embodiment of the invention;

FIG. 3B is a side view of the camera aperture of FIG. 3A;

FIG. 3C is an exploded diagram of a camera aperture

according to an embodiment of the invention;

FIG. 3D is a side view of the camera aperture of FIG. 3C;

FIG. 3E is an exploded view of a camera aperture in the

form of slits;

FIG. 3F is a view of an aperture offset from the optical axis;

FIG. 3G is a view of a camera aperture in the form of a

torus;

FIG. 4 is a diagram of a high frequency background pattern

according to an embodiment of the invention; and

FIG. 5 is a block diagram of a method for extracting a matte

according to one embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED

EMBODIMENT

System Structure

FIG. 1 shows a system 100 and method 500 for automatically

extracting a matte 141 from a video 110 acquired of a scene

120 according to an embodiment of our invention. The scene

120 includes a background (B) 121 and a foreground (F) 122. The scene can be a natural, real-world scene illuminated

only by ambient light.

The video 110 is acquired by a camera 101 including a

pinhole (narrow aperture) optical system 102 and a fore-

ground optical system 103. The optical systems 102-103 have a

single center of projection on an optical axis 160, and use a

beam splitter 151. The optical systems are calibrated with

respect to each other. The video 110 is provided to a processor

140 performing the method 500.

The video 110 includes sets of images 111-112 acquired

parallel in time. The set of images $I_p$ 111 is acquired at a large

depth-of-field 131, i.e., the images $I_p$ 111 are acquired with a

very narrow aperture focused at the foreground. The images

$I_w$ 111 can be approximated using a pinhole camera model, see

Equation (1). A corresponding set of wide aperture images $I_w$ 112 is acquired in parallel with a small depth-of-field 132

focused at the foreground.

FIG. 2 shows another embodiment of the invention. The

camera 201 uses a single optical system, and the images $I_p$

and $I_w$ of the sets 111-112 are serially interleaved in time.

Because the images are interleaved in time, pairs of corre-

sponding narrow aperture images $I_p$ and wide aperture

images $I_w$ may not be registered when the scene 120 includes

moving objects. In this case, a conventional optical flow pro-

cess can be used to register the sets of images $I_p$ and $I_w$.

Camera Aperture

Therefore, as shown in FIG. 3A, the camera 201 uses a fast

switching aperture 300. Most conventional camera apertures

use a mechanical shutter. The speed at which the mechanical

shutter can open and close is limited by the weight of the

leaves of the aperture and the strengths of the springs and

actuators driving the leaves of the aperture. Even if the aper-
ture mechanism were made very strongly, and driven with high-energy actuators, the resulting large mechanical motions would induce vibrations in the camera assembly blurring the image.

Most conventional camera apertures are mechanical, and include moving parts. There are two major problems with such apertures. First, the apertures are relatively slow to switch to different depths of field, and second, the rapid movement of the parts causes vibration in the camera body, which adds noise to the images, particularly if the imager is a CCD type of device. Therefore, it is desired to provide a fast switching camera aperture that operates on optical, and not mechanical, principals.

The aperture 300 can switch size at a frame rate of the camera 201 or higher, e.g., at a rate up to about 10 KHz. The aperture mechanism 300 includes two polarizers 301-303 having a first diameter D. The second polarizer 303 has a pinhole or narrow aperture (through hole) 304 having a second diameter d. The polarization of the polarizers 301-303 are rotated 90° with respect to each other, as shown. Therefore, light only passes through the pinhole 304.

A polarization rotator element 302, also having a diameter D, is disposed between the polarizers 301-303. The element 302 rotates the polarization of the light field passing through it by 90° when a voltage (V) is applied to the polarization rotator 302. For example, the element is a ferroelectric liquid crystal. Thus, when the voltage is applied to the element, the camera 201 has a large aperture diameter D because light passes though all three elements 301-303. Otherwise, absent the voltage, the light only passes through the pinhole 304 having an aperture diameter d. Commercial ferroelectric liquid crystal devices can switch from a zero-rotation state to a 90-degree rotation state in less than 10 microseconds with power inputs on the order of a few volts at a fraction of a milliamp. Other polarization rotators are possible to use. For example a Kerr cell can rotate polarization. Conventional nematic, supertwist liquid crystals can also rotate the polarization.

FIG. 3B is an end view of the aperture shown in FIG. 3A.

FIGS. 3C and 3D show another embodiment with multiple aperture settings possible. This aperture includes polarizers 321, 323, 325, and polarization rotators 322 and 324. Two though holes 331-332 are shown to provide three aperture settings. It should be understood that any number of fast switching aperture settings can be provided in this manner. The sizes of the apertures decrease in a direction of the light field passing through the aperture.

FIG. 3E shows a configuration where the through holes are slits 341-342 spaced apart at varying distances. This can be used to separate low (DC) and high (AC) frequency components in a light field. FIG. 3F shows an embodiment where the one or more through holes 351 are offset from the center (optical axis). FIG. 3G show an embodiment where the through hole 461 is a torus.

It should be noted, that if multiple through holes are used, it is actually possible to move the aperture along the optical axis, effectively changing the focal plane. This is not possible with mechanical apertures.

It should also be noted that various combinations of different through holes with differences in size, shape and offset from the optical axis can be used.

Method Operation

When the images are acquired according to the embodiments of the invention, then the following expression 502 (see FIG. 5) approximates the wide aperture images IP, 112:

\[ I_F = \alpha F + (1-\alpha) I_p, \]  
\[ B_F = B_{x} \times B_{y}, \]

and \( B_{x} \) is a point spread function (PSF).

Given a known background, we can directly solve for \( \alpha \). Although we begin with a known background, we avoid many of the drawbacks of difference matting by using two sets of images: the pinhole images 111 and the wide aperture images 112.

From Equations (1) and (2), we obtain an expression 503 for \( \alpha \):

\[ \alpha = C(I_F - I_p)/(B_{x} - B_{y}), \]

where \( C \) is a constant, e.g., 1.

FIG. 5 shows the steps of the basic method.

To produce better results, we use:

\[ \alpha = (I_F + I_p)/(B_{x} - B_{y}), \]

If color images are used, then \( B_{x}, B_{y}, I_p, \) and \( I_F \) are vectors. Thus, the expression for \( \alpha \) is

\[ \alpha = \|I_F + I_p\|/\|B_{x} - B_{y}\|, \]

where \( \|\| \) is a length operator for color vectors.

Given \( \alpha \) we can determine \( F \) using:

\[ \alpha F = I_F + (\alpha - 1) I_p \]

or

\[ \alpha F = I_F + (\alpha - 1) B \]

or

\[ \alpha F = 0.5x(I_F + I_p + (\alpha - 1)(B_{x} + B_{y})). \]

However, the \( \alpha \) expression can be ill-conditioned when \( B_{x} = B_{y} \). Therefore, alpha values for these pixels can not be determined reliably.

There are two possible solutions to this problem. A known background pattern can be used to guarantee that \( \|B_{x} - B_{y}\| \) is substantially larger than zero, see below. Alternatively, alpha values can be interpolated from the neighboring pixels. One can either use a threshold for the denominator \( \|B_{x} - B_{y}\| \) with the alpha values being interpolated from the neighbors when the denominator is less than a threshold, or one can use a confidence map for the value of the denominator. In this context, it is possible to use known scattered data interpolation methods, e.g., push-pull as described by Gortler et al., "The lumigraph," Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM Press, pp. 43-54, 1996, incorporated herein by reference.

Matte Background Pattern

In many applications, it is necessary to guarantee perfect matting results. In this setting, any type of incorrect matting and compositing is simply not acceptable. In such applications, the background can be illuminated using a projector, or the background can include a known pattern so that the expression used to determine \( a \) is well-conditioned.

According to Equation (6), the denominator \( \|B_{x} - B_{y}\| \) determines the precision of \( \alpha \). The background can have values
from $B_{\text{min}}$ to $B_{\text{max}}$. The image $B_{\text{p}}$ is a low-pass version of the background. An optimal low-pass (average) value of the background is:

$$B_{\text{p}} = 0.5x(B_{\text{min}} + B_{\text{max}}).$$ (10)

Optimal values for $B$ are $B_{\text{min}}$ and $B_{\text{max}}$. Thus, the value of the denominator is always $0.5x(B_{\text{max}} - B_{\text{min}})$. In the best case, when $B_{\text{min}} = 0$, the denominator is reduced to $0.5xB_{\text{max}}$ and one bit of alpha precision is lost, e.g., if $B_{\text{min}}$ is an 8 bit value, then the maximum precision of alpha is 7 bits.

There are many background patterns that give these optimal results. Here, we describe an example background. If the PSF $h_\theta$ is rotationally symmetric, then the pattern can be specified in 1D. Next, we determine allowed frequencies of the pattern.

If the size of the PSF $h_\theta$ is $k$ pixels, then the minimum frequency of the pattern is $k/0.5$ pixels. If the large depth-of-field pixels are aligned with the pattern, then the maximum frequency of the pattern is one pixel. In cases of higher frequencies, different patterns values are averaged. The use of color improves the conditioning of the problem when the pixels are misaligned with pattern transitions. It is desired to shift the pattern for different colors such that the value of the denominator is always large for at least one color.

For example, as shown in FIG. 4, a pattern in one dimension for the color red $401$ is shifted by $1/4$ in front of the pattern period with respect to the pattern $402$ for the blue color and the pattern $403$ for the green color. When the above patterns are superimposed and printed or projected onto a white surface in 2D, repeated vertical bars $410$ of the colors red, white, green/blue, and black are produced to produce the desired high frequency background pattern. In an actual implementation, the bars are about 2-3 millimeters wide. It should be understood that the pattern can easily be printed on wallpaper for covering an entire sound stage.

Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications may be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.

We claim:

1. A method for extracting an alpha matte from a video acquired of a scene, comprising the steps of:
   acquiring a set of narrow aperture images $I_{\beta}$ of a scene with a first camera aperture set to a relatively large depth-of-field and a narrow aperture, the scene including a background $B$ and a foreground $F$, and in which each narrow aperture image $I_{\beta}$ is expressed as $I_{\beta} = c_{\beta} + (1-\alpha)B$, where $c_{\beta}$ is an image in front of a background image $B$, and $\alpha$ is an alpha matte;
   acquiring a corresponding set of wide aperture images $I_{\beta}$ of the scene with a second camera aperture set to a relatively small depth-of-field and a wide aperture, and in which each wide aperture image $I_{\beta}$ is expressed as $I_{\beta} = c_{\beta} + (1-\alpha)B$, where $c_{\beta}$ is a point spread function; and
   extracting an alpha matte according to $\alpha = C - (c_{\beta} - I_{\beta})(B - B_{F})$, where $C$ is a constant.
2. The method of claim 1, in which $\alpha = (B - B_{F})(c_{\beta} - I_{\beta})(B - B_{F})$.
3. The method of claim 1, in which $C$ is equal to 1.
4. The method of claim 1, in which the narrow aperture images and the wide aperture images are color images, and the alpha matte is extracted according to $\alpha = (||B - B_{F}||I_{\beta} - I_{\beta})(B - B_{F})$, where $||\cdot||$ is a length operator for color vectors.
means for extracting an alpha matte according to \( \alpha = C + \frac{(I_f - I_p)}{(B - B_p)} \), where \( C \) is a constant.

18. A system for extracting an alpha matte from a video acquired of a scene, comprising:

5 an optical system configured to acquire a set of narrow aperture images \( I_p \) of a scene with a camera aperture set to a relatively large depth-of-field, the scene including a background \( B \) and a foreground \( F \), and in which each narrow aperture image \( I_p \) is expressed as \( I_p = \alpha F + (1 - \alpha) B \), where \( \alpha F \) is an image in front of a background image \( B \), and \( \alpha \) is an alpha matte, and further configured to acquire a corresponding set of wide aperture images \( I_w \) of the scene with the camera aperture set to a relatively small depth-of-field and a wide aperture, and in which

10 each wide aperture image \( I_w \) is expressed as \( I_w = \alpha F + (1 - \alpha) B \), where \( B_p = B(\otimes)h_p \), and \( h_p \) is a point spread function; and

means for extracting an alpha matte according to \( \alpha = C + \frac{(I_f - I_p)}{(B - B_p)} \), where \( C \) is a constant.

19. The system of claim 18, in which the camera aperture comprises:

a first polarizer;

a second polarizer with a polarization offset by 90° from the first polarizer, the second polarizer having a through hole; and

a polarizing rotator disposed between the first polarizer and the second polarizer.